All-In-Focus Synthetic Aperture Imaging

Tao Yang[1], Yanning Zhang[1], Jingyi Yu[2], Jing Li[3], Wenguang Ma[1], Xiaomin Tong[1], Rui Yu[4], and Lingyan Ran[1]

[1] SAIIP, School of Computer Science, Northwestern Polytechnical University, China
[2] Deptartment of CIS, University of Delaware, USA
[3] School of Telecommunications Engineering, Xidian University, China
[4] Deptartment of Computer Science, University College London, UK
yangtaonwpu@163.com, ynzhangnwpu@gmail.com, yu@eecis.udel.edu

Abstract. Heavy occlusions in cluttered scenes impose significant challenges to many computer vision applications. Recent light field imaging systems provide new see-through capabilities through synthetic aperture imaging (SAI) to overcome the occlusion problem. Existing synthetic aperture imaging methods, however, emulate focusing at a specific depth layer but is incapable of producing an all-in-focus see-through image. Alternative in-painting algorithms can generate visually plausible results but can not guarantee the correctness of the result. In this paper, we present a novel depth free all-in-focus SAI technique based on light-field visibility analysis. Specifically, we partition the scene into multiple visibility layers to directly deal with layer-wise occlusion and apply an optimization framework to propagate the visibility information between multiple layers. On each layer, visibility and optimal focus depth estimation is formulated as a multiple label energy minimization problem. The energy integrates the visibility mask from previous layers, multi-view intensity consistency, and depth smoothness constraint. We compare our method with the state-of-the-art solutions. Extensive experimental results with qualitative and quantitative analysis demonstrate the effectiveness and superiority of our approach.

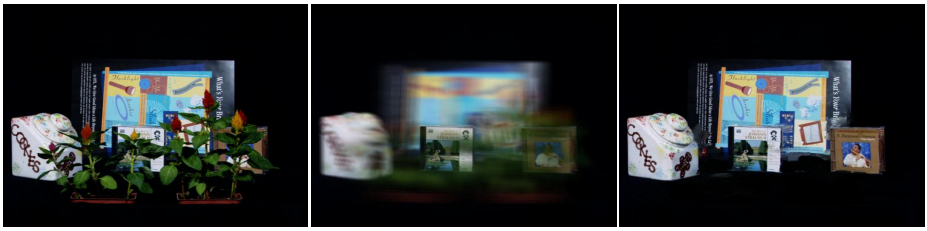
Keywords: occluded object imaging, all-in-focus synthetic aperture imaging, multiple layer visibility propagation.

1 Introduction

The capability of seeing through occlusions in heavily cluttered scenes is beneficial to many computer vision practical application fields, ranging from hidden object imaging to detection, tracking and recognition in surveillance. Since traditional imaging methods use a simple camera to acquire the 2D projection of the 3D world from a single viewpoint, they are unable to directly resolve the occlusion problem.

A fundamental solution to the problem is to exploit new imaging procedures. For example, emerging computational photography techniques based on generalized optics provide plausible solutions to capture additional visual information.

D. Fleet et al. (Eds.): ECCV 2014, Part VI, LNCS 8694, pp. 1–15, 2014.



(a) One camera view in the Stanford camera array

(b) Result of traditional synthetic image focused on the CD case

(c) Result of occluded object all-in-focus image by our method

Fig. 1. Comparison results of occluded object synthetic aperture imaging methods

In particular, Synthetic Aperture Imaging or SAI [1–15] provides a unique capability of seeing-through occlusions. SAI warps and integrates the multiple view images to simulate a virtual camera with an ultra-large convex lens and it can focus on different frontal-parallel [1] or oblique [2] planes with a narrow depth of field. As a result, objects lying on the virtual focus plane, even being occluded in reality, would be clearly imaged. A downside of traditional SAI, however, is that objects off the virtual focus plane would appear blurry even though they are not occluded. The see-through results, hence, depend on the depth of the virtual focus plane.

The objective of this work is to develop a novel algorithm to generate a depth-free all-in-focus image(as shown in Fig.1(c)). Here, the all-in-focus image would contain not only objects on the virtual focus plane of camera array, but also all objects observed inside the input scene at various depths. Depth free refers to that given a certain depth, the algorithm can see through all occluders in front of this depth and generate a clear and complete all-in-focus image of the scene contents behind it.

Different to in-painting algorithms [16, 17] which can generate visually plausible results but not guarantee the correctness of the result, our technique is based on the light field visibility analysis. For every 3D point, we trace all rays passing through it back to the camera array, and then construct a visibility layer in which the 3D point is visible in all active cameras. To recover the all-focus image behind a specific depth layer, we partition the scene into multiple visibility layers to directly deal with layer-wise occlusion, and apply an optimization framework to propagate the visibility information between multiple layers. On each layer, visibility and optimal focus depth estimation is formulated as a multiple label energy minimization problem. The energy integrates the visibility mask from previous layers, multi-view intensity consistency, and depth smoothness constraint. We compare our method with the state-of-the-art solutions on publica available Stanford and UCSD light field dataset, and a dataset captured by ourselves with multiple occluders. Extensive experimental results with qualitative and quantitative analysis demonstrate the superiority of our approach.

The organization of this paper is as follows. Section 2 introduces several related works. Section 3 presents the visibility layer propagation based imaging model. Section 4 details the visibility optimization algorithm. Section 5 describes

the dataset, implementation details and the experimental results. We conclude the paper and point out the future work in Section 6.

2 Related Work

Tremendous efforts have been made on developing light field imaging systems and post-processing algorithms. On the hardware front, light field camera arrays with different number of cameras, resolution, effective aperture size have been built, e.g., Stanford [3], CMU [4], UCSD [5], Alberta [6], Delaware [7], NPU [8], PiCam [15], etc., and the camera array synthetic aperture imaging technique has been proved to be a powerful way to see object through occlusion. Similar camera array technique has been adopted in producing movie special effects. For instance, in the 1999 movie The Matrix, a 1D camera array is used to create an impressive bullet dodging scene that freezes time but changes viewpoint towards the character.

On the algorithm front, one of the most important technique is synthetic aperture imaging (SAI). By integrating appropriate rays in the camera array, SAI can generate view that would be captured by a virtual camera having a large aperture. In addition, through shearing or warping the camera array images before performing this integration, SAI can focus on different planes in the scene. For example, the Stanford LF camera array by Levoy et al. [3] consists of 128 Firewire cameras, and for the first time align multiple cameras to a focus plane to approximate a camera with a very large aperture. The constructed synthetic aperture image has a shallow depth of field, so that objects off the focus plane disappear due to significant blur. This unique characteristic makes the synthetic aperture imaging a powerful tool for occluded object imaging.

Taking advantages of the geometry constraints of the dense camera array, Vaish et al.[11] present a convenient plane + parallax method for synthetic aperture imaging. A downside of their work, however, is that all rays from the camera array are directly integrated without further analysis. Thus, the clarity and contrast of their imaging result would be reduced by rays from the foreground occluders.

Visibility analysis through occlusion is a difficult but promising way to improve the occluded object imaging quality, and many algorithms have been developed in this way. Vaish et al.[12] study four cost functions, including color medians, entropy, focus and stereo for reconstructing occluded surface using synthetic apertures. Their method achieves encouraging result under slight occlusion; however the cost functions may fail under severe occlusion. Joshi et al.[10] propose a natural video matting algorithm using a camera array. Their method uses high frequencies present in natural scenes to compute mattes by creating a synthetic aperture image that is focused on the foreground object. Their result is inspiring and it has potential to be used for visibility analysis. However, this algorithm may fail in case of textureless background, and cannot deal with occluded object matting. Pei et al.[13] propose a background subtraction method for segmenting and removing foreground occluder before synthetic aperture imaging.

Their result is encouraging in simple static background, however since this approach is built on background subtraction, it cannot handle static occluder. In addition, their performance is very sensitive to cluttered background, and may fail under crowded scene.

Our method is perhaps closest to the work of Pei et al. [6] which solves the foreground segmentation problem through binary labelling via graph cuts. Instead of labelling the visibility and focusing depth, they label whether a point is on focus in a particular depth, and aggregate these focus labels in a given depth range to get a visibility mask for occluded object imaging. Although the result is encouraging,this method can only deal with front occluder(whose depth range need to be provided as a prior) labeling problem, and may fail if the occluder has severe self occlusion or there are multiple occluded objects due to lack of visibility propagation. In addition, the result of method [6] can only focus on particular depth of the scene instead of all-in-focus imaging, and the performance will be decreased in textureless background.

3 Visibility Layer Propagation Based Imaging Model

In this section we will introduce our multiple layer propagation based synthetic aperture imaging method. Instead of segmenting the observed scene into various depth layers, our approach segments the entire scene into multiple visibility layers. The **visibility layer** is defined on each layer as all the rays which are not occluded in any cameras, and computed by energy minimization. Points on each visibility layer do not necessarily need to correspond to the same object or surface.

By modelling the scene as visibility layers and propagating visibility information through layers, we can obtain the focussing depth and corresponded cameras for all the objects in the scene, including the occlusion object and occluded objects. So each visibility layer consists of pixels that are visible in all **active** cameras. The word active refers to the fact that the pixel position of the camera has not been labelled as occluded, e.g. not occupied by previous layers. Extraction of each visibility layer is based on the information of previous visibility layers. More precisely, according to occlusion mask information of previous layers, we firstly obtain the current visibility layer, then estimate the depth map of this layer, and finally update the occlusion mask.

For better understanding of the proposed method, we provide an example workflow with the Stanford Light Field data in Figure 2.

There are mainly two reasons why we introduce the concept of visibility layer. First, taking advantage of introduced visibility layer,occlusion problem can be tackled more directly. The visibility information is propagated from layer to layer, and in each layer occlusion mask needs to be updated only once. Second, segmenting the scene into visibility layers instead of depth layers is more beneficial as neighbouring pixels in the same layer tend to belong to the same object and depth smoothness constraint can be enforced when estimating the depth map.

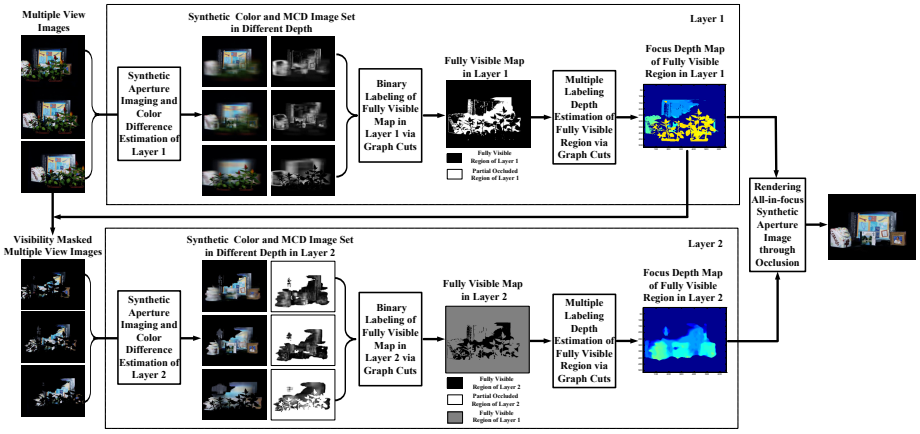


Fig. 2. Flowchart of the multiple layer visibility propagation based synthetic aperture imaging algorithm

Let L denotes the number of visibility layers in the scene. For each layer, we need to find a labelling function $f : \Omega \rightarrow \mathcal{L}$, where Ω refers to the set of all unmasked pixels in all images and $\mathcal{L} = \{0, d_1, d_2, \ldots, d_m\}$ denotes the set of possible labels of these pixels. $d_i \, (i = 1, 2, \ldots, m) > 0$ represents the depth range of our scene. For a pixel $\mathbf{x}$, if $f(\mathbf{x}) > 0$, then $\mathbf{x}$ is fully visible in all active camera views. Otherwise if $f(\mathbf{x}) = 0$, then $\mathbf{x}$ is partially occluded.

Considering the labelling redundancy of camera array(the labels in different cameras are highly related), the problem can be further simplified. Instead of labelling all the unmasked pixels of all the cameras, we label **all the pixels** of the reference camera equivalently(not only the unmasked pixels, as a masked pixel of the reference camera may still be fully visible in all the other active cameras). This means if there are N cameras in camera array, we only label all pixels of the reference camera view instead of labelling all the unmasked pixels of all cameras. Specifically, instead of finding above labelling function, we seek a more succinct labelling, $g : I_{ref} \rightarrow \mathcal{L}$, where I_{ref} refers to the whole image area of the reference camera. In our implementation, visibility and depth map is calculated first on the reference image, then the visibility and depth maps of all the other cameras are derived based on the calibration information of the camera array.

Therefore, for each layer ℓ , the problem of estimating fully visible pixels and corresponded depths can be formulated as a following energy minimization problem:

$$E(g; V_1, V_2, \ldots, V_{\ell-1}) = E_d(g) + E_s(g) \tag{1}$$

where the data term E_d is the sum of data cost of each pixel, and the smooth term E_s is a regularizer that encourages neighboring pixels to share the same label, while the visibility information $V_k (k = 1, 2, \ldots, \ell - 1)$ from previous layers is used to encode and block the occluded rays.

As the estimation of visibility information is coupled with depth value and can only be obtained by analyzing synthetic images of different depth layers, it is difficult to minimize energy function (1) directly. In this paper, we solve for $g = \{V_\ell, D_\ell\}$ by the following two optimization modules:1) optimize the visibility map V_ℓ in the reference camera, 2) calculate the depth map D_ℓ of visible pixels.

In order to obtain the fully visible map, in the first module we formulate this problem as binary energy minimization. If a pixel $\mathbf{x}$ is fully visible, it is labelled as 1, otherwise 0. Energy function consists of a unary data term which represents the cost of assigning a visibility label to a pixel and a pairwise smoothness term which accounts for smoothness prior of the visibility layer. This energy minimization problem is then optimized by graph cuts [18].

In the second optimization module, estimation of the optimal focus depth for pixels in each visible layer is formulated as a multiple label energy minimization problem and is also solved via graph cuts [18]. The energy function is composed of a unary data term which indicates the cost of assigning a depth label to a pixel, and a pairwise smoothness term which accounts for smoothness constraint of the depth map.

4 Multiple Layer Visibility Optimization

Since our method propagates the binary visibility map between multiple layers, for a certain layer $\ell \in \{1, 2, \ldots, L\}$, occluders in front of this layer have been labelled and can be easily removed in the images of all cameras. To make the notation uncluttered, we do not write previous visibility layers $V_k (k = 1, 2, \ldots, \ell - 1)$ explicitly unless necessary. As a result, the visibility energy function can be written as follows:

$$E\left(V_\ell\right) = E_d\left(V_\ell\right) + E_s\left(V_\ell\right) \tag{2}$$

Data Term: If a pixel is fully visible in current layer, it should be in focus for some depth value, and at the same time corresponding pixels that form the synthetic aperture image should be related by the same point of an object (except those occluded by previous layers). Since if a scene point is in focus, its corresponding pixel in the synthetic aperture image will have a good clarity and contrast,which can be measured by state-of-the-art focusing metrics. In addition, the corresponding pixels that form the synthetic aperture image should have a similar intensity value, which can be measured by various intensity constance metrics. In this paper, focusing metrics and intensity constance metrics are all referred to focusing metrics. We define the cost of labelling a pixel as fully visible based on its corresponding curve of focusing metrics in synthetic images of different depth layers.

The ideal curve of a fully visible pixel (Figure 3, point A) should satisfy the following two constraints: (1)it is unimodal throughout the focus depth scope, and (2) the curve reaches a global minimal, if and only if all visible rays intersect at the same point on an object in the scene. In contrast, a partially occluded pixel or a free point without focus should always have a large value through the entire focus depth scope (Figure 3, point C). That's because these points

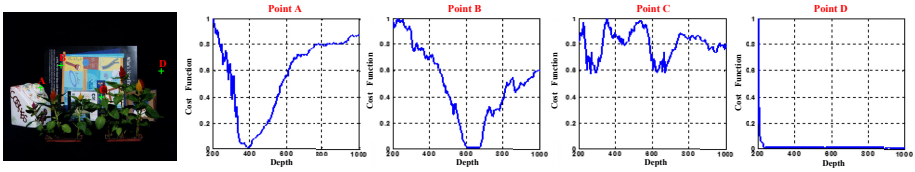


Fig. 3. Typical focusing curve of different kinds of points. Point A:fully visible texture region. Point B: Fully visible region with pure color. Point C: partial occluded region or free point. Point D: textureless region.

are only visible in some of the cameras, thus for unfocused depth and even for focused depth the cost of those points is high. A textureless object pixel should have a small value in a small range of depths around the focusing depth (Figure 3, point B),while a textureless background pixel should have a small value over a broad focus range due to its similarity with the neighborhood pixels (Figure 3, Point D). Besides, in Figure 3, Point D gives a sharp peak near the origin. That's caused by the position of focusing depth plane, when it's too close to the camera, the out of focus trouble results in an unexpected value.

Reasonably, we cannot estimate the depth of the textureless background pixels. Thus, according to the width of low value depth range we remove the textureless background region before our binary visibility optimization.

Based on the above analysis, we have compared different kinds of metrics to obtain the desired ideal curve. Part of the comparison result is shown in Figure 4.Figure 4(a) gives the input images of different cameras, while Figure 4(b), (c) and (d) display the synthetic aperture imaging result, variance image, and maximal color difference(MCD) image in different depths. Comparing Figure 4(c) and (d), we can see that our MCD Image could describe the minimal color difference more accurately than the viriance Image. Thus, the MCD measurement is more suitable for visibility analysis.

Figure 4(e) shows the corresponded curves of points A, B and C marked in Figure 4(a). The focus measures evaluated include DCT energy ratio (DCTR) [19], diagonal Laplacian (LAPD) [20], steerable filters (SFIL) [21], variance and MCD. For the first three focus measures,we compute the focus metric using a 5x5 pixel block on one hundred sampled focus planes. All the results are normalized and mapped to [0 1], where low value represents a good focus. The result indicates that for a point in textured region without occlusion, all focus measures can successfully find the focus point(point A in Figure 4(e)). However, when the textured point is occluded in some cameras(point B in Figure 4(e)), the curves of DCTR, LAPD, SFIL and variance measures are multimodal with multiple local minima. In contrast, MCD metric is more stable and more insensitive to occlusion. In the low texture region (Figure 4(e), point C), the first three measures contain many noises. In contrast, both the variance and MCD measure reach the global minimum around the ground truth. In addition, the MCD curve is more sharp than variance and more close to the ideal curve.

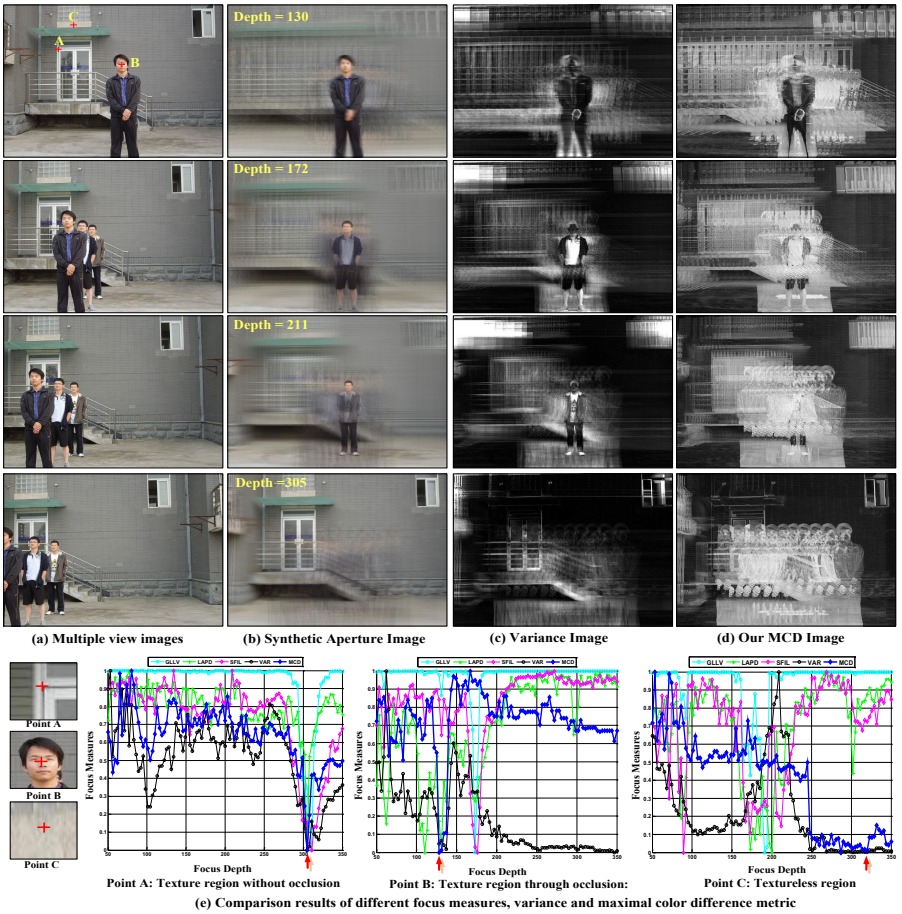


Fig. 4. Comparison results of focus measures for different kinds of points, the manually labeled ground truth focus depth is marked with the red arrow

Based on the analysis above, we select MCD measure to define the data cost $E_d(V_\ell)$ for each pixel $\mathbf{x}$ in the reference camera:

$$E_d(V_\ell) = \sum_{\mathbf{x} \in I_{ref}} \left(V_\ell(\mathbf{x}) - (1 - \min_{d \in \mathcal{D}}(MCD^d(\mathbf{x}))) \right) \tag{3}$$

where $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$ is the depth range of the scene, $MCD^d(\mathbf{x})(d \in \mathcal{D})$ is the MCD focus measure value of the pixel $\mathbf{x}$ in depth d:

$$MCD^d(\mathbf{x}) = \max_{\forall i \neq j} \left(|I_i^d(\mathbf{x}) - I_j^d(\mathbf{x})| \cdot B_i^\ell(\mathbf{x}) \cdot B_j^\ell(\mathbf{x}) \right) / 255 \tag{4}$$

$$B_i^\ell(\mathbf{x}) = \begin{cases} 0 \; if \sum_{\ell_0=1}^{\ell-1} V_{\ell_0}^i(\mathbf{x}) > 0 \\ 1 \quad otherwise \end{cases} \tag{5}$$

$I_i^d(\mathbf{x})$ represents the value of pixel $\mathbf{x}$ on the warped image of camera i in depth d. $B_i^\ell(\mathbf{x})$ is a binary map of camera i to mask fully visible pixels of previous layers. $V_{\ell_0}^i$ is the visibility layer ℓ_0 of camera i, and can be obtained easily from V_{ℓ_0} of the reference camera. If $B_i^\ell(\mathbf{x}) = 0$, $\mathbf{x}$ is occupied by previous layers, otherwise $B_i^\ell(\mathbf{x}) = 1$.

A good energy function should reach good solution when the energy is low. In order to achieve this, we design the data term of the visibility optimization model as Equation (3), which is introduced to classify all the pixels as visible or invisible. When min(MCD) is small, or data term is small, the probability that the point is occluded is low, thus the cost of assigning as a visible point is low. In addition, according to the definition of MCD, even if one of the camera view is occluded, the min(MCD) appears to be a large value, and the cost of assigning this point as a visible point is high by Equation (3). Thus for visibility labelling, it is straightforward to see that our data term should achieve its minimum when it is correctly assigned, and achieve a large value for occluded point, which is a perfect data term that we want.

Smoothness Term: The smoothness term $E_s(V_\ell)$ at layer ℓ is a prior regularizer that encourages overall labelling is smooth. The prior is that two neighbouring pixels have a higher probability to belong to the same object and should be both visible or occluded in the reference camera at the same time. Here we adopt the standard four-connected neighbourhood system, and penalize the fact if labels of two neighbouring pixels are different:

$$E_s(V_\ell) = \sum_{\substack{\mathbf{p}\in I_{ref} \\ \mathbf{q}\in\mathcal{N}\mathbf{p}}} S_{\mathbf{p},\mathbf{q}}(V_\ell(\mathbf{p}), V_\ell(\mathbf{q})) \tag{6}$$

$$S_{\mathbf{p},\mathbf{q}}(V_\ell(\mathbf{p}), V_\ell(\mathbf{q})) = \min(\tau_v, \beta(\mathbf{p},\mathbf{q}) \cdot |V_\ell(\mathbf{p}) - V_\ell(\mathbf{q})|) \tag{7}$$

$$\beta(\mathbf{p},\mathbf{q}) = h(|\min_{d\in\mathcal{D}}(MCD^d(\mathbf{p})) - \min_{d\in\mathcal{D}}(MCD^d(\mathbf{q}))|) \tag{8}$$

where τ_v and $\beta(\mathbf{p},\mathbf{q})$ denote the maximum and weight of smoothness term respectively. h is a decreasing weighting function that takes into account the MCD measure similarity between neighbouring pixels. The more similar MCD measure is, the weight will be higher and the smoothness constraint between pixels will be stronger.

In this paper, the parameters are given by experiment and we choose he inverse proportional function as $h(.)$. With the above data term and smoothness term, our energy function can be minimized via graph cuts [18].

After obtain V_ℓ, we formulate the optimal focus depth estimation inside the visible layer as a multiple label optimization problem, which is also solved via graph cuts in this work.

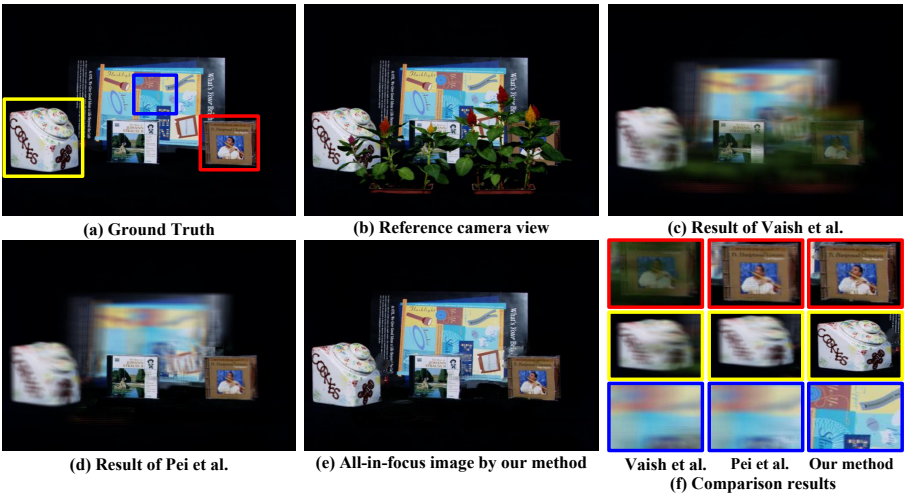


Fig. 5. Comparison result of different methods on CD case behind plants from Stanford

5 Experimental Results

We have compared the performance of our method with the synthetic aperture imaging methods of Vaish et al.[11] and Pei et al. [6] on four datasets, including the CD case behind plants from Stanford, the crowd surveillance scene from UCSD, and two dataset captured by ourselves. In addition, to illustrate that our method can be successfully applied when there are multiple visibility layers, we have captured another two datasets where there are multiple occluders.

To avoid explicit imaging for all the objects far away in the scene, we limit our search to a range of depths around the objects that our concern. For the CD case behind plants from Stanford and our own dataset, the accuracy of each method is compared with the ground truth separately. More implementation details are given below.

• Experiment 1: CD case behind plants

This dataset contains 105 views on a 21x5 grids (synthetic aperture size 60cm by 10cm) and the image resolution is 650x515. The scene contains some plants occluding two CD cases. Our goal is to estimate the depths for all the objects in the scene and image the scene behind the plants.

Figure 5 shows the comparison result of Vaish et al. [11], Pei et al. [6] and our method. We can see that all the three methods could see the occluded CD through the plants, as shown in Figure 5(c), (d) and (e). Pei et al. [6] is better than Vaish et al. [11] in imaging of the two CD cases. However, the objects away from the focus plane are blurring, including the CD on the right, who is just near the focus plane. All-in-focus image in Figure 5(e) shows much better clarity of our method and is closer to the ground truth in Figure 5(a).

Fig. 6. Comparison results on crowd surveillance data from UCSD light field data

Figure 5(f) gives the comparison of imaging for several local regions. It can seen that our method can give all-in-focus image for all the three objects of different depths, while the method of Vaish et al. [11], Pei et al. [6] can only focus on given depth plane.

Besides, we use the peak signal-to-noise ratio (PSNR) assessment to compare these methods quantitively (see Table 1). Calculation of PSNR is given in equation (13) and (14). The PSNR of our all-in-focus synthetic aperture image achieves 31.1088, which is much higher than 20.8774 of Pei et al. [6]and 18.1225 of Vanish et al [11].

$$PSNR = 10\log_{10}(I_{\max}^2/MSE) \tag{9}$$

$$MSE = \frac{1}{w \cdot h} \sum_{\mathbf{x} \in \mathcal{X}} \left(I(\mathbf{x}) - I'(\mathbf{x})\right)^2 \tag{10}$$

where w and h denote the image width and height, $\mathcal{X}$ is the image region, $I(\mathbf{x})$ is the pixel intensity value at $\mathbf{x}$ in the ground truth image and I' denotes the image to be assessed. $I_{\max} = 255$ is the maximal intensity value.

• Experiment 2: Crowd surveillance scene

The "Crowd" dataset is captured by UCSD with 8 synchronous views on an 8x1 grid. There are 276 frames and the image resolution is 640x480. The scene contains five people moving in the scene and they are frequently occluded by each other. Our goal is to see through the occluder in the front and image for all others continuously.

Figure 6 shows the comparison result of our method and Vaish et al [11]. In frame#210, the man in saffron cloth is occluded. In Vaish's result(Figure 6(b)), the occluded man is blurred by shadows from the occluder. In addition, people out of the focus plane are all blurred. In contrast, our approach could see

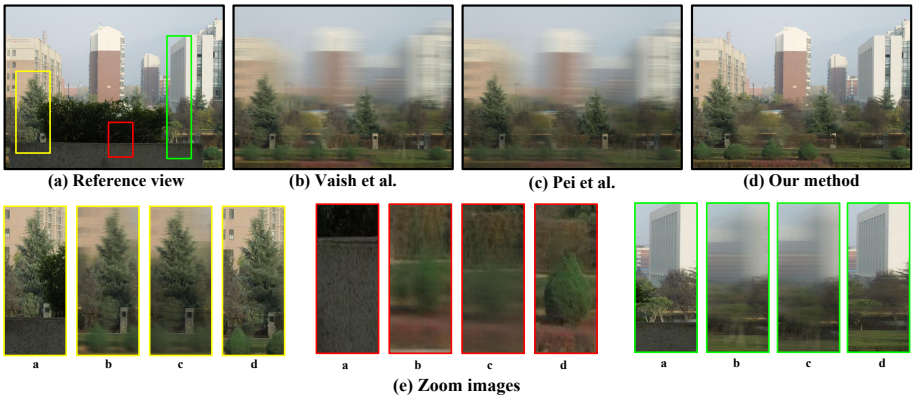


Fig. 7. Comparison results of different methods in the challenging outdoor scene

through occlusion and achieve a clear all-in-focus image (Figure 6(c)). Details of local region results are shown in Figure 6(d). Our method also shows better performance in frame#213 and #217 than Vanish's method.

The success of our work comes from the idea that for every synthetic aperture imaging result on each frame, the scene can be regarded as static and there are no moving objects. It's quite reasonable as no object would make obvious movements considering the high frequency that camera works. Figure 6 shows the result of several subsequent frames.

The limitation of our approach is that a scene point needs to be visible at least in two camera views, otherwise the black hole will appear in the all-in-focus image(Figure 6(c)).

• Experiment 3: Complex outdoor scene

To further test our method on severe occlusion cases, we have done another experiment with complex outdoor scene. As shown in Figure 7,the street ,trees and distant buildings are all occluded by nearby flowers. Our aim is to see the behind scene through the occlusion of front flowers. Comparison results of Vaish's method [11], Pei's method [6] and our method are shown in Figure 7(b) and Figure 7(c) and Figure 7(d). As Vaish's method only focus on a given depth plane and cannot eliminate front occluders completely, it cannot provide an all-in-focus image of the behind scene. And although Pei's method can remove some foreground occluder through foreground occluder segmentation and get a more clear result of target, the targets out of focus plane is still very blurring, for example the building shown in Figure 7(e). In comparison, our method could provide a depth free view point and all-in-focus image for any given depth range. For instance, Figure 7(d) shows the all-in-focus image of scene behind the flowers. Please note that although the depth change and occlusion in this scene is extremely complex, our method accurately gives the desired all-in-focus result.

• Experiment 4: Seeing through multiple occluded objects

Our method can be applied to the scene where there are multiple occluded objects. Due to visibility propagation between different layers, we can remove

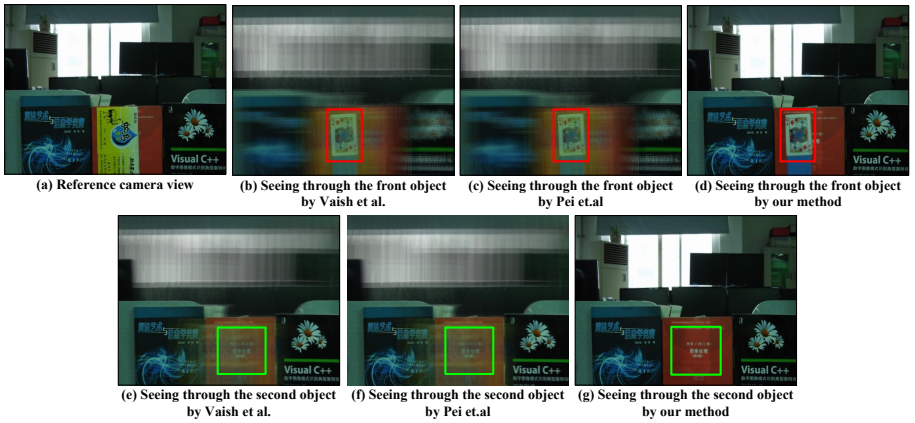


Fig. 8. Comparison results of synthetic aperture imaging through multiple objects

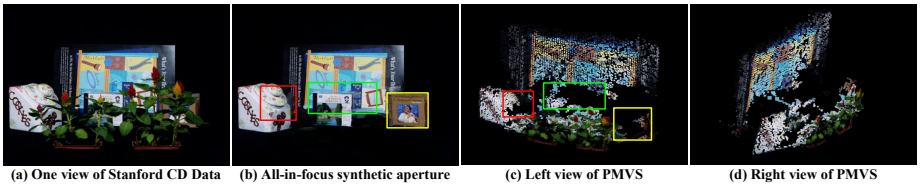


Fig. 9. Comparison with multi-view 3D reconstruction on CD case behind plants dataset

multiple occluders, focus on the occluded object and obtain an all-in-focus image of the occluded scene. Figure 8 shows the result of our synthetic aperture imaging method when there are multiple occluders. Figure 8(a)is the input image of the reference camera, it can be seen that the red book is occluded by the playing card, which is further occluded by front yellow box. The standard synthetic aperture imaging result of the playing card and red book is shown in Figure 8(b) and Figure 8(e)respectively. It can be seen that due to severe occlusion, Vanish's method[11] can only get a blurred image of the occluded object. The state-of-art synthetic aperture imaging result of playing card and red book is shown in Figure 8(c) and Figure 8(f) respectively. It can be seen that in the case of severe occlusion, Pei's method[6] can only get a blurred image of the occluded object due to the inaccuracy of the estimated foreground label. In comparison, our method can remove front occluders completely and provide an all-in-focus image of the scene behind the yellow box(Figure 8(d)) and even the playing card(Figure 8(g)).

• **Experiment 5: Comparison with multi-view 3D reconstruction**

Because it may be possible to apply stereo matching for producing see-through images, in this experiment we compare our approach with one of the state-of-the-art 3D reconstruction methods PMVS [22] on the public Stanford CD dataset. Two views of reconstruction result of the scene by PMVS are given in Figure

9(c) and (d). Due to the severe occlusion of foreground leaves and flowers, the reconstruction results of background CD contain many holes (as shown by green and yellow boxes). Our approach performs as an image-based depth peeling technique, it sequentially removes the front-most visible layers and generates an all-in-focus image of the observed scene through visibility layer prorogation(as shown in Figure 9(b)).

6 Conclusions

In this paper, we have presented a novel synthetic aperture imaging approach for creating all-in-focus images through occlusion. Different from existing synthetic aperture imaging algorithms, we have segment the scene into multiple visibility layers, and apply an optimization framework to propagate the visibility information between multiple layers to produce all-in-focus image even under occlusion.

We believe this approach is useful in challenging applications like surveillance of occluded people in crowded areas where seeing the people's appearance maybe of primary interest, or reconstructing hidden objects through severe occlusion, or even rendering a depth free viewpoint image. In the future, we would like to design more robust cost functions for the focus depth estimation, and extend our work to unstructured light field imaging through occlusion with handhold mobile phone.

Acknowledgements. We thank the anonymous reviewers for their valuable comments and detailed suggestions. This work is supported by the National Natural Science Foundation of China (No.61272288,No.61231016), NSF grants IIS-CAREER-0845268 and IIS-1218156, Foundation of China Scholarship Council (No.201206965020, No.201303070083), Foundation of NPU New Soaring Star, NPU New People and Direction (No.13GH014604), Graduate Starting Seed Fund of NPU,and NPU Soaring Star (No.12GH0311).

References

1. Isaksen, A., McMillan, L., Gortler, S.J.: Dynamically reparameterrized light fields. In: SIGGRAPH, pp. 297–306 (2000)
2. Vaish, V., Garg, G., Talvala, E., Antunez, E., Wilburn, B., Horowitz, M., Levoy, M.: Synthetic aperture focusing using a shear-warp factorization of the viewing transform. In: CVPR, pp. 129–134 (2005)
3. Wilburn, B., Joshi, N., Vaish, V., Talvala, V.E., Antunez, E., Barth, A., Adam, A., Horowitz, M., Levoy, M.: High performance imaging using large camera arrays. ACM T GRAPHIC 24(3), 765–776 (2005)
4. Zhang, C., Chen, T.: A self-reconfigurable camera array. In: SIGGRAPH (2004)
5. Joshi, N., Avidan, S., Matusik, W., Kriegman, D.J.: Synthetic aperture tracking: tracking through occlusions. In: ICCV, pp. 1–8 (2007)
6. Pei, Z., Zhang, Y.N., Chen, X., Yang, Y.H.: Synthetic aperture imaging using pixel labelling via energy minimization. Pattern Recognition 46(1), 174–187 (2013)

7. Ding, Y.Y., Li, F., Ji, Y., Yu, J.Y.: Synthetic aperture tracking: tracking through occlusions. In: ICCV, pp. 1–8 (2007)
8. Yang, T., Zhang, Y.N., Tong, X.M., Zhang, X.Q., Yu, R.: A new hybrid synthetic aperture imaging model for tracking and seeing people through occlusion. IEEE TCSVT 23(9), 1461–1475 (2013)
9. Basha, T., Avidan, S., Hornung, A., Matusik, W.: Structure and motion from scene registration. In: CVPR, pp. 1426–1433 (2012)
10. Joshi, N., Matusik, W., Avidan, S.: Structure and motion from scene registration. In: SIGGRAPH, pp. 779–786 (2006)
11. Vaish, V., Wilburn, B., Joshi, N., Levoy, M.: Using plane + parallax for calibrating dense camera arrays. In: CVPR, pp. 2–9 (2004)
12. Vaish, V., Szeliski, R., Zitnick, C.L., Kang, S.B., Levoy, M.: Reconstructing occluded surfaces using synthetic apertures: stereo, focus and robust methods. In: CVPR, pp. 2331–2338 (2006)
13. Pei, Z., Zhang, Y.N., Yang, T., Zhang, X.W., Yang, Y.H.: A novel multi-object detection method in complex scene using synthetic aperture imaging. Pattern Recognition 45(4), 1637–1658 (2011)
14. Davis, A., Levoy, M., Durand, F.: Unstructured light fields. EUROGRAPHICS 31(2), 2331–2338 (2012)
15. Venkataraman, K., Lelescu, D., Duparr, J., McMahon, A., Molina, G., Chatterjee, P., Mullis, R.: Picam: An ultra-thin high performance monolithic camera array. ACM T. Graphics 32(5), 1–13 (2013)
16. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: SIGGRAPH, pp. 417–424 (2000)
17. Tauber, Z., Li, Z.N., Drew, M.S.: Review and preview: disocclusion by inpainting for image-based rendering. IEEE TSMC 37(4), 527–540 (2007)
18. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. TPAMI 26(9), 1124–1137 (2004)
19. Lee, S.Y., Yoo, J.T., Kumar, Y., Kim, S.W.: Reduced energy-ratio measure for robust autofocusing in digital camera. Signal Processing Letters 16(2), 133–136 (2009)
20. Thelen, A., Frey, S., Hirsch, S., Hering, P.: Interpolation improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood-size, and height value interpolation. TIP 18(1), 151–157 (2009)
21. Minhas, R., Mohammed, A.A., Wu, Q.M.J., Sid-Ahmed, M.A.: 3D shape from focus and depth map computation using steerable filters. In: Kamel, M., Campilho, A. (eds.) ICIAR 2009. LNCS, vol. 5627, pp. 573–583. Springer, Heidelberg (2009)
22. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multi-view stereopsis. TPAMI 32(8), 1362–1376 (2010)